Endogenous giant viruses shape intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*

Mohammad Moniruzzaman*[1] and Frank O. Aylward*[1]

[1] Department of Biological Sciences, Virginia Tech, Blacksburg, VA
Email address for correspondence: monir@vt.edu, faylward@vt.edu

Abstract:

Chlamydomonas reinhardtii is an important eukaryotic alga that has been studied as a model organism for decades. Despite extensive history as a model system, phylogenetic and genetic characteristics of viruses infecting this alga have remained elusive. We analyzed high-throughput genome sequence data of numerous *C. reinhardtii* isolates, and in six strains we discovered endogenous genomes of giant viruses reaching over several hundred kilobases in length. In addition, we have also discovered the entire genome of a closely related giant virus that is endogenized within the genome of *Chlamydomonas incerta*, one of the closest sequenced phylogenetic relatives of *C. reinhardtii*. Endogenous giant viruses add hundreds of new gene families to the host strains, highlighting their contribution to the pangenome dynamics and inter-strain genomic variability of *C. reinhardtii*. Our findings suggest that endogenization of giant viruses can have profound implications in shaping the population dynamics and ecology of protists in the environment.

Introduction:

Giant viruses that infect diverse eukaryotes have recently emerged as a widely distributed viral group in the biosphere (1, 2). These viruses belong to a broadly defined group that has been recently classified as the phylum *Nucleocytoviricota* (3). Although giant viruses are primarily studied as agents of mortality of diverse protists and metazoans, recent studies have demonstrated that they can drastically shape the genomes of their hosts through endogenization (4). Specifically, Giant Endogenous Viral Elements (GEVEs) that are up to several thousand kilobases long have been identified in diverse green algae (4), providing compelling evidence that endogenization of diverse giant viruses can profoundly influence host genome evolution and alter the evolutionary trajectory of host eukaryotes.

Chlamydomonas reinhardtii is a widely used model photosynthetic eukaryote with history as a model organism dating back to the 1950s (5). Here, we report that endogenous giant viruses are common in field isolates of *C. reinhardtii*. Through identification of near-complete genomes of giant viruses endogenized in field isolates, we show that *C. reinhardtii* is a host of diverse giant viruses from distinct phylogenetic affiliations. Our results establish this widely-studied green alga as a model to study the mechanistic and evolutionary aspects of giant virus endogenization in diverse eukaryotic lineages, and provides the first insights on the genomic complexity and phylogenetic history of viruses infecting *C. reinhardtii* in nature.

Results:

We analyzed publicly available high-throughput genome sequencing data for 33 wild strains of *C. reinhardtii*. This data was originally generated for population genomic studies of diverse *C. reinhardtii* strains (6–8). After assembly and annotation (see Methods for details), we identified GEVEs in six of the wild strains (Figure 1B). In five of these (CC-2936, 2937, 2938, 3268, and GB-66), the GEVEs range from 315-356 Kb in size and harbored all but one *Nucleocytoviricota* hallmark genes, indicating that near-complete genomes of endogenous giant viruses have been retained in these strains (Figure 1B, Dataset S1). In contrast, CC-3061 harbors a GEVE ~113 Kb in size with 5 out of the 10 hallmark genes, indicating that part of the GEVE was lost over the course of evolution (Supplementary Methods, Dataset S1). We also analyzed the assembled genome of *Chlamydomonas incerta,* a species phylogenetically closest to *C. reinhardtii*, for which a long-read assembled genome has been recently reported (9). This analysis revealed a GEVE ~475 Kb long which is integrated within a single 592 Kb contig of *C. incerta*. (Figure 1B)

The GC-content of the *C. reinhardtii* GEVEs ranged from 58.27% (CC-2938) to 60.72% (CC-3268), close to the overall genomic GC content of *C. reinhardtii* (64%) (10). Similarly, the GC content of the *C. incerta* GEVE was 64.8%, resembling the overall GC content of the *C. incerta* genome (66%) (9) (Figure 1B). The GEVEs also contained numerous spliceosomal introns, ranging from 25 (CC-3061) to 72 (*C. incerta*) which has been previously found to be a feature of GEVEs present in other members of the Chlorophyta (4). Together, these results suggest the GEVEs were subjected to GC-content amelioration and intron-invasion since endogenization. In addition, the GEVE in *C. incerta* was flanked by highly repetitive regions on both ends (Figure 2A). The repetitive region at the 5'-end harbors several reverse transcriptases and transposases (Dataset S1). These regions also have higher intron density compared to the GEVE region itself, and lower number of Giant Virus Orthologous Group (GVOG) hits consistent with their eukaryotic provenance (Figure 2A). This suggests that near-complete genomes of giant viruses can integrate within highly repetitive regions of eukaryotic genomes, potentially with the facilitation of transposable elements.

Using a newly established taxonomy of *Nucleocytoviricota* (11), we determined the phylogenetic position of the *C. reinhardtii* and *C. incerta* GEVEs and their relationships with other chlorophyte GEVEs that were recently reported (4) (Figure 1A). Five of the strains harbored GEVEs that formed a cluster within the *Imitervirales* order, consistent with their high average amino acid identity. The GEVE in *C. incerta* was the closest phylogenetic relative of the *Imitevirales* GEVEs in *C. reinhardtii,* indicating that closely-related giant viruses infect closely related *Chlamydomonas* species in nature. These GEVEs formed a sister clade with the GEVEs present in six other volvocine algae and belonged to the *Imitevirales* family 12 (Figure 1A). These results suggest that many viruses infecting volvocine algae in nature are closely related, and provides a phylogenetic framework for future efforts on detection and isolation of viruses infecting these algal lineages. In contrast to the GEVEs that could be classified as *Imitervirales*, the GEVE in CC-2938 strain belonged to the *Algavirales* (Figure 1A), demonstrating that *C. reinhardtii* is infected by multiple phylogenetically distinct lineages of giant viruses.

The GEVEs in *C. reinhardtii* encoded 99 (CC3061) to 254 (CC2937) genes, while the *C. incerta* GEVE encoded 355 genes. Most of the genes were shared among the *Imitervirales C. reinhardtii* GEVEs, consistent with their high average amino acid identity (AAI) to each other (Figure 1C, D). These GEVEs also shared a high number of orthogroups with the *C. incerta* GEVE. In contrast, only a few orthogroups were shared between the *Imitevirales and the Algavirales* GEVEs consistent with the large phylogenetic distance between these lineages. Between ~44-55% of the genes in the *C. reinhardtii* and *C. incerta* GEVEs have matches to Giant Virus Orthologous Groups (GVOGs), confirming their origin in the *Nucleocytoviricota* (Figure 1B). In addition, different genes in these regions have best matches to giant viruses, bacteria, and eukaryotes, which is a common feature of *Nucleocytoviricota* members given the diverse phylogenetic origin of the genes in these viruses (12) (Figure 2A). Based on the Cluster of Orthologous Group (COG) annotations, a high proportion of the GEVE genes are involved in transcription, and DNA replication and repair; however, genes encoding translation, nucleotide metabolism and transport, signal transduction, and posttranslational modification were also present, consistent with the diverse functional potential encoded by numerous *Nucleocytoviricota* (Figure 2B).

A previous study has shown that several field strains of *C. reinhardtii* harbor many genes that are absent in the reference genome (7), which were possibly acquired from diverse sources. To quantify the amount of novel genetic material contributed by giant viruses to *C. reinhardtii,* we estimated the number of unique gene families in the analyzed *C. reinhardtii* field strains that are absent in the reference strain CC-503. On average ~2.24% of the genes in the field strains were unique compared to the reference strain (Figure 2C). Moreover, the GEVE-harboring field strains have significantly more unique genes compared to those without GEVEs, (Two-sided Man-Whitney U-test p-value <0.05). These results suggest that endogenization of giant viruses is an important component of inter-strain variability in *C. reinhardtii*. Recent studies have highlighted the importance of horizontal gene transfer (HGT) in structuring the pangenome of diverse eukaryotes (13, 14), and genes originating from endogenous *Nucleocytoviricota* were found to shape the genomes of many algal lineages, including members of the Chlorophyta (4, 15). Compared to the GEVE-free strains, GEVE-containing strains harbored a significantly higher proportion of genes from several COG categories including Transcription, Signal Transduction, and Replication and Repair (Two sided Mann-Whitney U test p-value <0.05) (Figure 2C). All together, these GEVEs contributed many genes with known functions, including glycosyltransferases, proteins involved in DNA repair, oxidative stress, and heat shock regulation (Dataset S1).

Discussion:

While much work remains to elucidate the role of GEVEs in shaping the ecological dynamics of *C. reinhardtii*, several possibilities remain open. Some of the genes contributed by the GEVEs can be potentially co-opted by the host, leading to changes in certain phenotypes compared to closely related strains without GEVEs. Strain-specific endogenization can also potentially lead to intraspecific variations in chromosome structure, partly mediated by the GEVE-encoded mobile elements (16). Finally, it is also possible that some of these GEVE-loci can produce siRNAs that might participate in antiviral defense, and similar phenomena has been suggested for the virus-like loci in the genome of moss (*Physcomitrella patens*) (17). Recent studies on large-scale endogenization of giant virus genomes in diverse green algae and patchwork of viral genes in many algal lineages suggest that interactions between giant viruses and their algal hosts frequently shape the host genomes (4, 15), and therefore represents a major research frontier for studying the effect of giant viruses in influencing the (HGT) landscape. Our results indicate that endogenization of giant viruses can lead to large-scale genomic variability not only between algal species, but also between strains within the same population. Results reported in this study therefore represent an important advance in our understanding of how giant viruses shape the genome evolution of their hosts, while also expanding the scope of *C. reinhardtii* as a model organism to study the evolutionary fate and consequences of giant virus endogenization.

Methods:

All methods and relevant citations are available in the ‘Supplementary Information’ file.

Data and Code availability:

Dataset S1 contains information regarding the raw data source, GEVE functional annotations, hallmark gene distribution in each GEVE and coverage information of the partial GEVE in CC-3061. Dataset S2 contains figures related to the Method section - TNF dendrogram for the CC-2938 GEVE and promer alignment plots between *C. reinhardtii Imitevirales* GEVEs and *C. incerta* GEVE.

All the GEVE fasta files, unique gene fasta in each of the strains and their annotations, and concatenated alignment file used to build the phylogenetic tree in Figure 1 are available in Zenodo: https://zenodo.org/record/4958215

Code and instructions for ViralRecall v2.0 and NCLDV marker search scripts are available at: github.com/faylward.

Acknowledgement

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This work was supported by grants from the Institute for Critical Technology and Applied Science and the NSF (IIBR-1918271) and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A.

Conflict of interest statement

The authors declare no conflict of interest relevant to the content of the manuscript.

Figure legends:

Figure 1: General features and phylogeny of the GEVEs. A) Maximum likelihood phylogenetic tree of the GEVEs and representative members from diverse NCLDV families constructed from a concatenated alignment of seven NCLDV hallmark genes (see Methods). Individual families within each order are indicated with abbreviations (IM - Imitevirale, AG - Algavirales) followed by family numbers, as specified in Aylward et al, 2019 (11). IDs of the GEVEs are indicated in bold-italic. **B)** Basic statistics of the GEVEs present in various field strains of *C. reinhardtii* and the GEVE present in the C. *incerta* genome. **C)** Heatmap representing the number of orthologous groups of proteins shared across all the GEVEs. **D)** Heatmap representing the average amino acid (AAI) % among the GEVEs.
***** Length includes the eukaryotic regions flanking the *C. incerta* GEVE.

Figure 2: GEVE genomic and functional characteristics. A) Circular plots of two representative GEVEs in *C. reinhardtii* and the GEVE present in *C. incerta.* For *C. reinhardtii* one representative *Imitevirales* GEVE (CC-2937) and the Algavirales GEVE (CC-2938) are shown. Circle plots show Giant Virus Orthologous Group (GVOG) hidden Markov model (HMM) hits, spliceosomal introns and the best LAST hit matches (see Supplementary Methods). Internal blue links delineate the duplicated regions. The eukaryotic regions flanking the *C. incerta* GEVE are delineated with light blue stripes. **B)** Functional potential encoded in the GEVEs as EggNOG categories. Proportion of genes in each category is normalized across all the NOG categories except category S (function unknown) and R (general function prediction). Raw functional annotations are present in Dataset S1. **C)** Unique genes in the field strains of *C. reinhardtii* compared to the reference strain CC-503. The heatmap represents % of unique genes that can be classified in different EggNOG categories (except category [R] - genel function prediction and [S] - function unknown). Categories marked with '**' are significantly overrepresented in the GEVE-containing strains compared to those without GEVEs. The bar plot on top of the heatmap represents % of unique genes in each strain. The names of the abbreviated categories are the same as in figure 2B. GEVE-containing strains have significantly higher percentages of unique genes compared to the strains without GEVEs.

References:

1. M. Moniruzzaman, C. A. Martinez-Gutierrez, A. R. Weinheimer, F. O. Aylward, Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1710 (2020).

2. H. Endo, *et al.*, Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* **4**, 1639–1649 (2020).

3. E. V. Koonin, *et al.*, Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84** (2020).

4. M. Moniruzzaman, A. R. Weinheimer, C. A. Martinez-Gutierrez, F. O. Aylward, Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* (2020) https:/doi.org/10.1038/s41586-020-2924-2.

5. S. Sasso, H. Stibor, M. Mittag, A. R. Grossman, From molecular manipulation of domesticated to survival in nature. *Elife* **7** (2018).

6. R. J. Craig, *et al.*, Patterns of population structure and complex haplotype sharing among field isolates of the green alga Chlamydomonas reinhardtii. *Mol. Ecol.* **28**, 3977–3993 (2019).

7. J. M. Flowers, *et al.*, Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas reinhardtii. *Plant Cell* **27**, 2353–2369 (2015).

8. A. R. Hasan, J. K. Duggal, R. W. Ness, Consequences of recombination for the evolution of the mating type locus in Chlamydomonas reinhardtii. *New Phytol.* **224**, 1339–1348 (2019).

9. R. J. Craig, A. R. Hasan, R. W. Ness, P. D. Keightley, Comparative genomics of Chlamydomonas. *Plant Cell* (2021) https:/doi.org/10.1093/plcell/koab026.

10. S. S. Merchant, *et al.*, The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007).

11. F. O. Aylward, M. Moniruzzaman, A. D. Ha, E. V. Koonin, A Phylogenomic Framework for Charting the Diversity and Evolution of Giant Viruses. *PloS Biol.* 19(10):e3001430.

12. J. Filée, N. Pouget, M. Chandler, Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol. Biol.* **8**, 320 (2008).

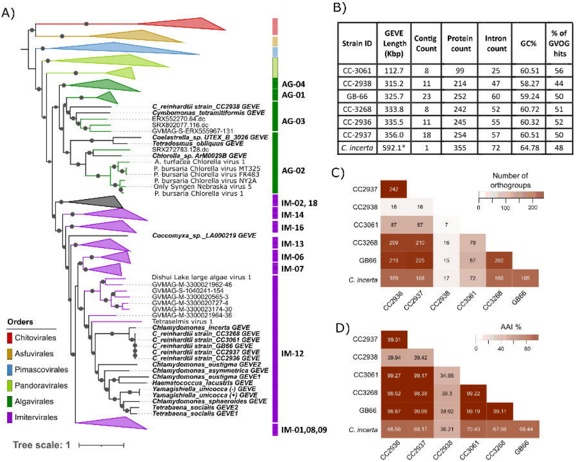
13. X. Fan, *et al.*, Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. *Sci Adv* **6**, eaba0111 (2020).

14. S. J. Sibbald, L. Eme, J. M. Archibald, A. J. Roger, Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends Parasitol.* **36**, 927–941 (2020).

15. D. R. Nelson, *et al.*, Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. *Cell Host Microbe* **29**, 250–266.e8 (2021).

16. J. Filée, Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. *Curr. Opin. Virol.* **33**, 81–88 (2018).

17. D. Lang, *et al.*, The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).



B)

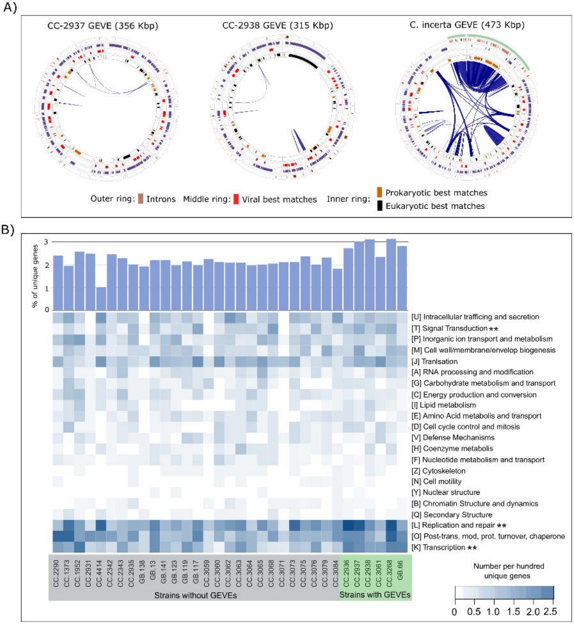
Strain ID	GEVE Length (Kbp)	Contig Count	Protein count	Intron count	GC%	% of GVOG hits
CC-3061	112.7	8	99	25	60.51	56
CC-2938	315.2	11	214	47	58.27	44
GB-66	325.7	23	252	60	59.24	50
CC-3268	333.8	8	242	52	60.72	51
CC-2936	335.5	11	245	55	60.32	52
CC-2937	356.0	18	254	57	60.51	50
C. incerta	592.1*	1	355	72	64.78	48

C)

	CC2936	CC2937	CC2938	CC3061	CC3268	GB66
CC2937	242					
CC2938	16	16				
CC3061	87	87	7			
CC3268	209	210	16	79		
GB66	219	225	15	87	202	
C. incerta	169	168	17	72	166	165

D)

	CC2936	CC2937	CC2938	CC3061	CC3268	GB66
CC2937	99.31					
CC2938	39.94	39.42				
CC3061	99.27	99.17	34.55			
CC3268	98.52	98.39	39.3	99.22		
GB66	98.87	99.09	38.02	99.19	99.11	
C. incerta	68.56	68.17	36.21	70.43	67.66	68.44


A)
CC-2937 GEVE (356 Kbp)
CC-2938 GEVE (315 Kbp)
C. incerta GEVE (473 Kbp)
Outer ring: Introns
Middle ring: Viral best matches
Inner ring: Prokaryotic best matches
Eukaryotic best matches
B)
% of unique genes
0
1
2
3
[U] Intracellular trafficing and secretion
[T] Signal Transduction **
[P] Inorganic ion transport and metabolism
[M] Cell wall/membrane/envelop biogenesis
[J] Tranlsation
[A] RNA processing and modification
[G] Carbohydrate metabolism and transport
[C] Energy production and conversion
[I] Lipid metabolism
[E] Amino Acid metabolis and transport
[D] Cell cycle control and mitosis
[V] Defense Mechanisms
[H] Coenzyme metabolis
[F] Nucleotide metabolism and transport
[Z] Cytoskeleton
[N] Cell motility
[Y] Nuclear structure
[B] Chromatin Structure and dynamics
[Q] Secondary Structure
[L] Replication and repair **
[O] Post-trans. mod, prot. turnover, chaperone
[K] Transcription **
CC.2290
CC.1373
CC.1952
CC.2931
CC.4414
CC.2342
CC.2343
CC.2935
GB.138
GB.13
GB.141
GB.123
GB.119
GB.117
CC.3059
CC.3060
CC.3062
CC.3063
CC.3064
CC.3065
CC.3068
CC.3071
CC.3073
CC.3075
CC.3076
CC.3079
CC.3084
CC.2936
CC.2937
CC.2938
CC.3061
CC.3268
GB.66
Strains without GEVEs
Strains with GEVEs
Number per hundred unique genes
0
0.5
1.0
1.5
2.0
2.5